Constructing a Lexicon of Dutch Discourse Connectives

Bourgonje, Peter; Hoek, Jet; Evers-Vermeul, Jacqueline; Redeker, Gisela; Sanders, Ted; Stede, Manfred

Published in:
Computational Linguistics in the Netherlands Journal

Computational Linguistics in the Netherlands Journal 8 (2018) 163–175 Submitted 06/2018; Published 12/2018

Constructing a Lexicon of Dutch Discourse Connectives

Peter Bourgonje[*] BOURGONJE@UNI-POTSDAM.DE
Jet Hoek[**] J.HOEK@UU.NL
Jacqueline Evers-Vermeul[**] J.EVERS@UU.NL
Gisela Redeker[***] G.REDEKER@RUG.NL
Ted Sanders[**] T.J.M.SANDERS@UU.NL
Manfred Stede[*] STEDE@UNI-POTSDAM.DE

[*] *Universität Potsdam, Germany*

[**] *Universiteit Utrecht, The Netherlands*

[***] *Rijksuniversiteit Groningen, The Netherlands*

Abstract

We present a lexicon of Dutch Discourse Connectives (DisCoDict). Its content was obtained using a two-step process, in which we first exploited a parallel corpus and a German seed lexicon, and then manually evaluated the candidate entries against existing connective resources for Dutch, using these resources to complete our lexicon. We compared connective definitions in the research traditions of the two languages and accommodated the differences in our final lexicon. The DisCoDict lexicon is made publicly available, both human- and machine-readable, and targeted at practical use cases in the domain of automatic discourse parsing. It also supports manual investigations of discourse structure and its lexical signals.

1. Introduction

A central task in discourse processing is inferring coherence relations. These relations connect pieces of text and express a certain sense, such as Contrast, Elaboration or Cause. The words or phrases that explicate these kinds of relations are referred to as *connectives* or *discourse connectives*, which belong to the broader category of *discourse markers* (Knott and Dale 1994, Redeker 1991). Connectives do not form a recognized (closed) word class, but constitute a syntactically heterogeneous group, including conjunctions, different kinds of adverbials, and prepositions. Like other discourse markers (Degand et al. 2013), connectives are often multifunctional linguistic expressions. First, they sometimes also serve a role connecting elements within a sentence; compare (1), in which *and* has a sentential reading, and (2), in which *and* has a discourse reading. Not all researchers consider such sentential cases instances of discourse connective use (see also Section 5). Second, connectives can represent multiple different senses. While *and* in (2) marks a Conjunction relation, *and* in (3) marks a Contrast relation.

(1) *Sophie was tired and happy.*

(2) *Sophie was tired and she wanted to go to sleep.*

(3) *Sophie was tired, and Jonathan was wide awake.*

Having a lexicon of connectives for a specific language can be a very helpful resource for language learners, providing them with instructions on which semantic relations can be expressed with which surface forms. In addition, such a resource is an essential starting point of many approaches to automatic discourse parsing. Connective lexicons have already been developed for several other

languages: DiMLex for German (Stede 2002), LICo for Italian (Feltracco et al. 2016), LexConn for French (Roze et al. 2012) and LDM-PT for Portuguese (Mendes and Lejeune 2016).

In this paper we describe the process of creating a connective lexicon in both human- and machine-readable form for Dutch. The resulting lexicon is made publicly available under the name DisCoDict.[1] In Section 2, we first provide some more background on the German DiMLex lexicon, which served as the starting point for creating the Dutch lexicon, and on the Penn Discourse Treebank, which we have used for attributing sense labels to the connectives. Section 3 describes our approach toward populating a list of candidate entries and the steps taken to filter these candidates, and specifies how the different attributes (relating to sense, syntactic category, etc.) were attached to the entries. Section 4 compares the outcome of this step to already existing connective resources for Dutch. Section 5 discusses the challenges of reconciling a practical approach, targeted at a lexicon useful for application scenarios, with a theoretical grounding of the entries in such a lexicon, and why certain entries are included or excluded from it. Finally, Section 6 provides an overview of what the entries look like, and Section 7 sums up the conclusions and provides suggestions for further improvement.

2. Background: DiMLex and Penn Discourse Treebank

This section provides background information on the initial discourse marker lexicon (for German) that inspired the format of DisCoDict, and on the large English-language corpus annotation project that generated the sense label classification we have used for annotating the connectives in the lexicon: the Penn Discourse Treebank (PDTB).

2.1 DiMLex

As a starting point for creating a dictionary of Dutch discourse connectives, we have used DiMLex (Stede 2002, Stede and Umbach 1998), a German lexicon that aims at exhaustively listing all connectives for German. While the first version covered around 150 frequent connectives, the latest version (Stede and Neumann 2014) contains 275 entries and is considered as "by and large complete." From the outset, DiMLex was devised as a multi-purpose resource, which could be used for

- supporting the manual annotation of connectives in text with a suitable annotation tool;
- informing programs performing automatic discourse parsing, e.g., following the frameworks of Rhetorical Structure Theory (Mann and Thompson 1988) or PDTB (see Section 2.2);
- informing text generation programs that need to include connectives in order to signal semantic or pragmatic relations.

The definition for selecting entries in DiMLex was adapted from the pioneering work of Pasch et al. (2003) on German, who state that a lexical item X is a connective when:

- X is not inflectable,
- X does not assign case to its syntactic environment,
- X expresses some specific, two-place semantic relation,
- the arguments of the relational meaning of X are propositional structures,
- the verbalizations of the arguments of the relational meaning of X can be clauses.

1. `https://github.com/discourse-lab/DisCoDict`

For DiMLex, it was decided to drop the second requirement, as it rules out prepositions. While for Pasch et al. (2003) this served as a syntactic filter to reduce the set of candidate terms, the perspective of DiMLex was to account for items like *wegen* 'due to' and *trotz* 'despite', because in many cases it is possible to paraphrase a causal or concessive clause with an NP headed by these prepositions (e.g., *the concert was canceled because it rained* vs. *the concert was canceled due to the rain*).

Note that the definition does not restrict the notion of 'connective' to be a single word. In fact, DiMLex distinguishes between different kinds of phrasal units, specifying whether the different parts have to be contiguous (e.g., *anstatt dass* 'instead of' followed by a clause) or separated (e.g., *wenn - dann* 'if - then'). The important constraint for phrasal expressions to be included is their being frozen: they cannot be inflected and do not allow for inserting modifying material. This delineates the border to the less-constrained so-called 'secondary connectives' (Danlos et al. 2018), which DiMLex does not account for.

The idea to build a lexical resource for connectives is also motivated by the fact that these items can take part in standard lexical relations such as synonymy (e.g., German *obschon* 'although' vs. *obzwar* 'although'), plesionymy (or near-synonymy; e.g., English *although* and *though* differ in formality), antonymy (e.g., *if* vs. *unless*), hyponymy (e.g., the general *but* can serve the purposes of various more specific contrastive and concessive markers). Further, as remarked in the introduction, many connectives are polysemous, as they can signal various relations (e.g., *since* for temporal or causal relations). The following examples from (Stede 2002, p. 110) illustrate the range of possible paraphrases for Concession (in the RST sense), and hence the set of items that a connective lexicon should include for this relation:

(4) *We were in SoHo;* {*nevertheless* | *nonetheless* | *however* | *still* | *yet*}*, we found a cheap bar.*

(5) *We were in SoHo, but we found a cheap bar anyway.*

(6) {*Despite* | *Notwithstanding*} *the fact that we were in SoHo, we found a cheap bar.*

(7) {*Although* | *Even though*} *we were in SoHo, we found a cheap bar.*

The task addressed by DiMLex was to account for such connections as much as possible, and to provide the syntactic, semantic, and pragmatic features that distinguish similar connectives from one another. Over time, however, breadth of coverage was considered more important than "depth", and therefore the latest release of DiMLex has many more connectives than the original version, but the descriptions are not as detailed as originally envisaged. Still, for many purposes such as discourse parsing, the basic syntactic description and the PDTB sense information form a useful basis; these also form the backbone of the new DisCoDict.

Specifically, the XML format that DisCoDict borrows from DiMLex (an example entry will be shown in Section 6) specifies orthographic variants of the connective;[2] whether it can have a non-connective reading; its discourse sense in terms of the Penn Discourse TreeBank (see below); its syntactic category (subordinating conjunction, coordinating conjunction, adverbial, preposition); information on linear ordering of the linked material; and example sentences.

2.2 Penn Discourse Treebank (PDTB)

As the name suggests, the PDTB (Prasad et al. 2008) is not a lexicon but a discourse-annotated corpus, which is built on top of the syntactically-annotated Penn Treebank (Marcus et al. 1994). In contrast to the DiMLex approach, corpus annotation began without any given list of connectives; annotators had to identify candidate items themselves, and then also mark the two spans (or 'arguments') that are being connected. This step is executed sentence-by-sentence in every text. The

2. In German, this is quite relevant due to 'offical' changes in spelling in 1996; but we also include some uncommon spellings as they are sometimes used in social media.

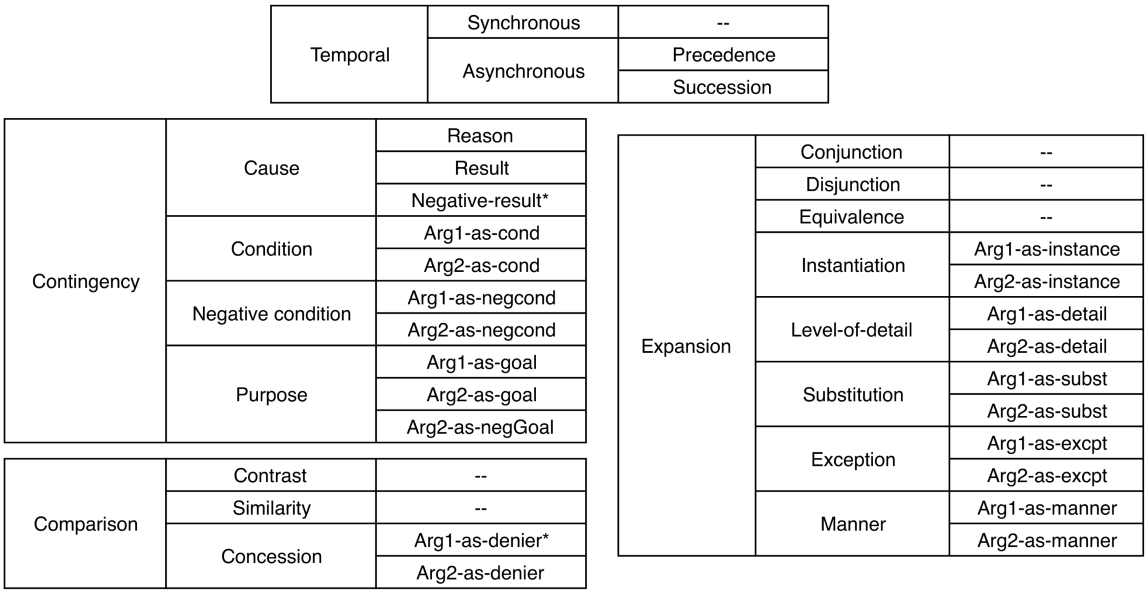

Figure 1: The PDTB-3 Sense Hierarchy.

underlying assumption is that there must be some kind of semantic/pragmatic connection between adjacent sentences and clauses, even if it is not explicitly signaled by a connective (or some other phrasal expression). This is why the PDTB project also annotated *implicit* relations between adjacent material. Importantly, the same inventory of relations is used for explicit and implicit cases. The whole corpus consists of approximately 40k relations, of which ca. 18k are explicit and ca. 16k are implicit (the remaining cases being distributed over alternative lexicalization, entity relation and no relation cases, see Prasad et al. (2008) for details). In the explicit relations, 101 unique connectives have been identified by the annotators.[3] The relations, also called *senses* of the connectives, have evolved a little over the course of the PDTB project, and since they are now empirically well-tested, we also adopt the inventory for DisCoDict.

The set of relations in the current version (PDTB-3) is organized as a taxonomy with three layers, where the first layer broadly distinguishes between Temporal, Comparison, Contingency, and Expansion relations. We show the taxonomy in Figure 1. For a slightly different earlier version (PDTB-2), there is an extensive annotation manual (PDTB Research Group 2007); the new PDTB-3 version is introduced in Webber et al. (2016).

3. Method

The first design decision was to adopt the basic XML format of DiMLex (as described in Section 2.1) for our Dutch lexicon; one reason for this choice is that it has already been adapted in recent years to several other languages (as mentioned in Section 1), indicating a sufficient degree of language-neutrality. Then, the central methodological steps for our work were the exploitation of a parallel corpus for gathering an initial set of Dutch candidate connectives; the mechanics of deciding on the items to be included in the lexicon; the sense assignment procedure; and the definition of additional attributes. In the following we discuss these steps in turn.

3. These in turn have served as the basis for an English connective lexicon; see Das et al. (2018).

3.1 Parallel corpus exploitation

Apart from the structure of DiMLex as a basis for our lexicon, we also used its German entries and exploited a parallel German-Dutch corpus to speed up the population of the initial lexicon of Dutch discourse connectives. We expected this parallel corpus approach, also explored specifically in the context of discourse connectives by for instance Cartoni et al. (2013) and Versley (2010), to be faster than manual translation of the German entries. In addition, we expected to find a broad range of entries and spelling variations using this data-driven approach.

We used two parallel corpora: the German-Dutch section of Europarl (Koehn 2005), containing 44M words, and News-Commentary11 (Tiedemann 2012), containing 492K words. We extracted word alignments from these sentence-aligned corpora using MGIZA++ (Gao and Vogel 2008). Then, using the German entries as a seed lexicon, we looked up their alignments on the Dutch side, and collected the alignment frequencies of the corresponding words and phrases. Note that the lookup procedure did not differ structurally between words and phrases. In both cases, single words (stand-alone or in a phrase) could correspond to zero, one or multiple target words. The target representation was collected in a key-value structure, where the key is the position in the sentence and the value is the word. This list was then sorted by position to return the target word or phrase, which is potentially discontinuous.

An obvious drawback of our data-driven approach, however, is that it is sensitive to both the quality of the seed lexicon and the domain of the parallel corpus used. We selected DiMLex, because it is a lexicon developed over the course of several years and can reasonably be expected to be stable and exhaustive.[4] Moreover, the syntactic similarities between German and Dutch are likely to result in relatively high quality word alignments, thereby finding more relevant words and phrases automatically. To counteract the domain impact, and more generally to obtain a lexicon as complete as possible, we compared the output of this bilingual lookup to existing Dutch resources (see Section 4.1).

3.2 Initial lexicon population

Because the word alignments are not guaranteed to be correct, we discarded any alignment with a frequency of 1% or lower to filter for unlikely translations; if a particular word or phrase in German aligned to the word or phrase in Dutch in only 1% or less of its absolute frequency in the German corpus, it was discarded. In addition to incorrect word alignments, there was another source of errors: ambiguity. If an entry on the German seed lexicon is used in its non-connective reading, it may be aligned to an element on the Dutch side that cannot be a connective at all. Since at the time of writing we did not have a connective classifier readily available for the German language,[5] we did not have an automated way of dealing with this potential error source.

The entries remaining after this automatic filtering were judged manually. Usually, a German seed contained several alignments that scored above the threshold in automatic filtering. The correct Dutch alignment was part of this set, but was often accompanied by several irrelevant alignments. A first selection that discarded these irrelevant entries could be made relatively easily on the basis of roughly the following four groups:

- Cases where the relevant Dutch connective was found, but preceded by punctuation marks. For example, the result set for *gemäß* 'according to' included both the relevant entry *overeenkomstig* and an irrelevant duplicate *, overeenkomstig* (preceded by a comma). In addition, several German connectives were aligned to only a comma. This error is probably due to instances where the coherence relation was left implicit, resulting in a misalignment.

- Cases where the seed connective and/or its most prototypical Dutch equivalent is a phrase, and the alignment would be only part of that phrase. Examples include *umso mehr* 'even

4. Though the latter remains also a matter of connective definition, as discussed in Section 5.
5. However, such a classifier is under construction, see Bourgonje and Stede (2018).

more' aligned to *te meer* (part of *des te meer*) and *anlässlich* 'because of/owing to' aligned to *naar aanleiding* 'in response' (as part of the phrase *naar aanleiding (daar)van* 'in response to (this)').

- Cases where the alignment could be vaguely associated with the connective, but made no sense in isolation. Examples are *zuerst* 'first' aligned to *in*, a preposition that could be part of the connective *in de eerste plaats* 'in the first place', but should not be included in the lexicon separately; and *soweit* 'as far as' being aligned to *begrepen* 'understood'. A common phrase attached to *soweit* in German is *Soweit ich es verstanden habe* 'as far as I have understood it', which would translate to *Voor zover ik het begrepen heb*, which could explain the alignment.

- Cases where the seed connective had such a low frequency that different variations were less likely to be filtered out by the 1% threshold. This included *währenddessen* 'meanwhile' aligned to *terwijl in slaan waarbij geblokkeerde vermijden* 'lit. while in hitting whereby blocked avoid', *dat reeds hachelijke levensomstandigheden* 'that already precarious living conditions' and *evident dat ondertussen* 'evident that in the meantime'. Because these cases were typically quite long, a simple heuristic filtering out candidates containing more than five words resulted in perfect precision (i.e., filtering out only garbage and no relevant entries).

The remaining entries were checked further by three native Dutch speakers (the first three authors of this paper), leaving a set of 157 entries resulting from this semi-automatic method to populate the lexicon.

3.3 Sense assignment

The Dutch lexicon uses the same set of senses as the German lexicon and is based on PDTB 3.0 (see Section 2.2). Unfortunately we have no annotated Dutch corpus using the same set of sense labels available to establish the possible sense or senses of a particular connective. In order to assign a sense to a connective, we therefore looked at its closest German counterpart and took the sense label of this entry in DiMLex, assigning multiple senses where applicable. In case the sense deviated from its closest German counterpart (i.e., the Dutch connective could express fewer or more different senses), a human evaluator (native Dutch speaker with extensive knowledge of coherence relations) assigned the relevant sense or senses as appropriate; all senses were then checked by two other human evaluators (also native Dutch speakers with extensive knowledge of coherence relations). A case in point is the German connective *anlässlich* 'because of/owing to', which in addition to the Result sense has the senses Reason and Synchronous, whereas its closest Dutch counterpart (*naar aanleiding van*) only has the Reason sense.

3.4 Additional attributes

In addition to listing the connective and its sense(s), DiMLex provides several additional attributes that we also specified for our Dutch lexicon. The syntactic labels attached to the connectives are based on the syntactic categories also present in the German lexicon, and are one of the following: `adv` for adverbials, `cco` for co-ordinating conjunctions, `csu` for sub-ordinating conjunctions, `prep` for prepositions and `other` for remaining cases. The value of this label was decided upon by a human coder, supported by the syntactic label of the closest German counterpart of the connective and the part-of-speech label attached to it by the Alpino parser (Van Noord 2006) when parsing the connective's example sentence.

Attributes that specify the options for the ordering of the arguments are also provided. According to the PDTB, arg2 is "the argument that appears in the clause that is syntactically bound to the connective" and arg1 is "the other argument" (p.1, PDTB Research Group 2007). The attributes in the lexicon that specify the argument order make explicit whether arg1 can appear before or after (or both before and after) arg2, and whether or not arg2 can be inserted in arg1.

Every entry comes with an example sentence of the connective in its discourse reading and, if applicable, an example sentence of the entry in its non-connective reading (e.g. *echter* 'but' versus *echter* 'more real').

Finally we specify whether or not the connective can take two finite clauses as its arguments, to accommodate the difference in definitions observed in the Dutch and the German research tradition regarding connectives (see Section 5). An example entry of the final lexicon is included in Section 6.

4. Validation

After semi-automatically compiling the lexicon content using the seed lexicon and parallel corpora, we compared the initial lexicon with 157 entries this method resulted in to existing inventories of connectives to (i) assess the effectiveness of the method described in Section 3 for generating a complete inventory of Dutch connectives and (ii) identify and supplement missing entries in order to improve the completeness of the DisCoDict lexicon.

4.1 Comparing the generated lexicon to other connective inventories

While there is no other connective lexicon for Dutch, there are some resources that list Dutch connectives. We selected three such resources to compare the result of the semi-automatic extraction method to. The first list we consulted was compiled by Van Wijk and Kempen (1980). Since this list contains all types of Dutch function words, we selected only words from categories 2 (coordinating and subordinating conjunctions) and 7 (sentence connecting adverbs). In addition, we consulted the connective list put forward by Pander Maat (2002) and the (non-exhaustive) Dutch connective list generated by means of manual translation spotting in Hoek et al. (2017) and Hoek (2018).[6] After removing duplicate entries and a few non-connective words (mostly stance markers), the three lists together yielded a set of 137 unique Dutch connectives.[7] 87 of the connectives from this list were also included in the semi-automatically generated list constituting the first draft version of the DisCoDict lexicon, which means that 55% of the DisCoDict entries also occurred in the list, while 64% of the list items also occurred in the first draft version of the DisCoDict lexicon.

The comparison of the largely automatically generated lexicon that was the starting point of DisCoDict to the list generated on the basis of other Dutch connective inventories illustrates both the strengths and the weaknesses of the parallel corpus lookup approach. The parallel corpus method did not yield an exhaustive Dutch lexicon, missing many connectives, including some fairly frequent, prototypical connectives such as *doordat* 'because (of that),' *toen* 'then,' or *ook al* 'though.' On the other hand, the approach identified more connectives than it missed, and identified many connectives that were not included in existing inventories of Dutch connectives, which mostly focused on single-word expressions. Another benefit is that for the connectives it does identify, the approach also generates syntactic and sense label information that otherwise has to be supplemented by hand.

4.2 Supplementing the lexicon

We supplemented the DisCoDict lexicon with connectives that were included in the combined list of Dutch connectives but not in the initial version of the lexicon. We searched for the connectives in the Europarl corpus to find representative examples to include in the lexicon, and to find any non-connective uses of the entries. Sense labels for the new entries were supplied by a human coder (native Dutch speaker with extensive knowledge of coherence relations), using Europarl examples to help guide decisions. All labels were then checked by two other coders.

Seven connectives from the compiled list did not occur (as connectives) in the Europarl corpus, which comfirmed our intuition that they were largely archaic. These connectives (*desniettegen-*

6. Hoek's point of departure was a set of eight English connectives, which yielded a list of 63 Dutch connectives.
7. This number does not include the seven archaic entries, see Section 4.2.

staande, *dewijl*, *eerdat*, *hierenboven*, *mitsdien*, *naardien*, *aangemerkt dat*) were left out of the lexicon. The current version of the DisCoDict lexicon consists of 207 entries. Table 1 gives a schematic overview of the lexicon throughout the validation process.

Automatically generated lexicon		**Comparison lexicon and connective list**			**Final lexicon**
n = 157	→	+lexicon +list	n = 87	→	n = 207
		+lexicon -list	n = 70		
		-lexicon +list	**n = 50**		

Table 1: Evolution of the lexicon.

5. Theoretical vs. practical considerations

As reflected by the choice for a standardized format (XML) and a sense hierarchy (PDTB) often used in recent approaches to automatic discourse parsing (cf. Biran and McKeown 2015, Lin et al. 2014, Oepen et al. 2016, Wang and Lan 2015) (making it easier to compare performance to other systems and languages), an important part of the work done is targeted at practical usability. As already mentioned in Section 1, one example of a use case is discourse parsing, where a comprehensive and exhaustive list of words and phrases that can signal a discourse relation is a useful resource. Another implementation example comes from an online multilingual connective database; connective-lex,[8] to which our lexicon is uploaded. This database, described in more detail in Scheffler et al. (2018), is targeted at multilingual research on connectives specifically, allowing queries for connectives with the same sense or syntactic category (or combinations thereof) for various languages. In the database, the lexicons for all individual languages can be consulted for monolingual applications as well.

At the same time, we have strived to offer a lexicon of connectives that has solid theoretical foundations. One challenging aspect here is that while many definitions of word groups are based on syntactic constraints, connectives form a syntactically heterogeneous group. Since we took DiMLex as the starting point for populating our lexicon of Dutch connectives, we originally also adopted its definition of 'connective' (see Section 2). However, the definition of 'connective' can differ between frameworks, and indeed there appear to be some differences between the way in which this notion is operationalized in DiMLex and the way in which it is commonly defined within Dutch approaches to discourse coherence. First, a crucial point of departure is that while connective literature originating from the Netherlands often assumes coherence relations to hold between segments that are minimally clauses (e.g., Evers-Vermeul 2005, Sanders et al. 1992, Sanders and van Wijk 1996), or "constituent discourse units" in the RST-based study of Van der Vliet and Redeker (2014), this is not a restriction set in DiMLex, which includes certain nominalized arguments for prepositional connectives such as *aufgrund* 'by virtue of' and *wegen* 'because of'. Because these connectives cannot take a clausal complement, their Dutch counterparts *krachtens* and *vanwege* would not be considered to be connectives in most Dutch approaches to coherence relations. A second point of departure concerns the inclusion of both single words and multi-word expressions in the DiMLex lexicon, while most Dutch coherence researchers preserve the category of connectives for single-word markers of coherence relations (a.o. Evers-Vermeul 2005, Sanders et al. 1992), using the term 'cue phrases' for a broader category that includes both single-word and multi-word markers of coherence relations (compare Knott and Dale 1994).

Concerning the second difference, DisCoDict followed the more liberal definition of DiMLex, including both single-word connectives and multi-word cue phrases. To account for the first difference, we added an additional column to the lexicon which contains information about whether or not the entry can connect two clauses. This solution accommodates a stricter definition of 'connec-

8. `http://connective-lex.info/`

tive' without losing any information from the generated lexicon. In addition, we made sure that for each connective that can connect two finite clauses, the coherence relation in the connective example holds between two finite clauses; these examples would be considered as coherence relations in all approaches. Finally, we created distinct entries for connectives that cannot connect two finite clauses by itself, but for which this does become an option when *dat* 'that' is added. Consider *behalve* 'except', which can connect two finite clauses, and *behalve that* 'except that', which cannot.

6. Summary

The final lexicon contains 207 entries, of which 94 (45%) are adjectives, 48 (23%) are subordinating conjunctions, 13 (6%) are co-ordinating conjunctions, 33 (16%) are prepositions and 19 (9%) have the category `other` assigned to them. There are a total of 30 different senses for all entries, and 21 entries (10%) have more than one sense. 86 entries (42%) have both a connective and a non-connective reading. The number of entries that can connect two finite clauses is 159 (77%) and the remaining 48 (23%) entries cannot connect two finite clauses, but necessarily take a non-finite clause, a prepositional phrase or a noun phrase as one of its arguments. The XML structure of the lexicon is based on DiMLex (see Section 3), with the only deviation in structure being the addition of a `finiteClauseArg` node, indicating whether the connective takes two finite clauses (value of 1) or not (value of 0). One example entry is shown in Listing 1.

Listing 1: Example entry of DisCoDict.

```
<entry id="c27" word="daarbuiten">
    <orths>
      <orth type="cont" canonical="1" onr="c27o1">
        <part type="single">daarbuiten</part>
      </orth>
      <orth type="cont" canonical="0" onr="c27o2">
        <part type="single">Daarbuiten</part>
      </orth>
    </orths>
    <non_conn_reading>
      <example>Wij weten dat deze strijd tegen de corruptie in de hele Europese
          Unie en ook daarbuiten moet worden gevoerd.</example>
    </non_conn_reading>
    <ambiguity>
      <non_conn>1</non_conn>
      <sem_ambiguity>0</sem_ambiguity>
    </ambiguity>
    <stts>
      <example>Natuurlijk moeten de Verenigde Staten, ook militair, in staat zijn
          om in hun eigen achtertuin de vrede te bewaren. Daarbuiten moeten we
          vooral doen waar we goed in zijn: dat is praten, praten, nog eens
          praten, compromissen sluiten en uiteindelijk betalen.</example>
    </stts>
    <syn>
      <cat>adv</cat>
      <integr/>
      <ordering>
        <ante>0</ante>
        <post>1</post>
        <insert>0</insert>
      </ordering>
      <sem>
        <pdtb3_relation sense="exception-arg1-as-except"/>
      </sem>
    </syn>
  <finiteClauseArg>1</finiteClauseArg>
</entry>
```

Similar to DiMLex, sentence-initial and all-lowercase variants are included for all entries by default under orthographic variants. This is extended where necessary (for example, the often abbreviated form of *dat wil zeggen*: *d.w.z.*, is added as an orthographic variant). The type of an orthographic variant specifies whether the variant is continuous (`cont`) or discontinuous (`discont`). Continuous variants consist of one part, of which in turn the type is either `single` for single tokens or `phrasal` for multiple tokens. Discontinuous variants consist of multiple parts, for which again the types are either `single` or `phrasal`. For example, the entry for *zowel...als* includes the orthographic variant *zowel...als ook*, which is discontinuous and consists of two parts. The first part, *zowel*, is of the type `single` and the second part, *als ook* is of the type `phrasal`. There are 142 single-word entries (69%) and 65 multi-word entries (31%). Of the multi-word entries, 6 are discontinuous and the remaining 59 are continuous.

The `non_conn_reading` node is empty if the entry only has a connective reading, and includes an example otherwise (as in Listing 1). This `non_conn` and sense ambiguity (`sem_ambiguity`) is also made explicit in the `ambiguity` node with a 1 for ambiguous or 0 for non-ambiguous cases. The `stts` node contains a usage example. The `syn` node contains subnodes for part-of-speech, argument ordering and sense.

7. Conclusion

We have presented a Dutch connective lexicon that is both human- and machine-readable, and targeted at practical use cases. Whether or not this lexicon can be considered exhaustive, is inherently connected to the definition of discourse connective. We have addressed differences in German and Dutch research traditions in this area, and used an additional attribute for entries in our lexicon to accomodate both interpretations. To generate the lexicon, we first exploited two parallel German-Dutch corpora, using a German connective lexicon (DiMLex) as a seed lexicon to obtain candidate entries through word alignments. The resulting candidate entries were then compared to existing resources for Dutch and the lexicon was completed upon using these resources. The first stage of this process offered speedy population of a list of candidates, and allowed exploitation of semantic and syntactic information from the source lexicon. We have shown, however, that it lacks in coverage and made up for this in the second stage, exploiting existing Dutch resources. The lexicon can aid in scenarios ranging from (human) language learning to machine translation and discourse parsing. The uploading of DisCoDict in the multilingual database connective-lex being one practical example, we consider the evaluation of this lexicon in other use cases, such as shallow discourse parsing, an important piece of future work.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this paper. Bourgonje's contribution was enabled by the project 'Anaphoricity in Connectives', funded by the Deutsche Forschungsgemeinschaft (DFG). Evers-Vermeul's contribution was enabled by a grant awarded by the Executive Board of Utrecht University to the AnnCor project, work package Discourse Annotation.

References

Biran, Or and Kathleen McKeown (2015), PDTB discourse parsing as a tagging task: The two taggers approach, *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2015)*, Association for Computational Linguistics, pp. 96–104.

Bourgonje, Peter and Manfred Stede (2018), Identifying explicit discourse connectives in German, *Proceedings of the 19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2018)*, Association for Computational Linguistics, Melbourne, Australia, pp. 327–331.

Cartoni, Bruno, Sandrine Zufferey, and Thomas Meyer (2013), Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique, *Dialogue & Discourse* **4**, pp. 65–86.

Danlos, Laurence, Katerina Rysova, Magdalena Rysova, and Manfred Stede (2018), Primary and secondary discourse connectives: definitions and lexicons, *Dialogue & Discourse* **9** (1), pp. 50–78.

Das, Debopam, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede (2018), Constructing a lexicon of english discourse connectives, *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Melbourne, Australia.

Degand, Liesbeth, Bert Cornillie, and Paola Pietrandrea (2013), Discourse markers and modal particles: Two sides of the same coin?, *in* Degand, L., B. Cornillie, and P. Pietrandrea, editors, *Discourse markers and modal particles: Categorization and description*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 1–18.

Evers-Vermeul, Jacqueline (2005), *The development of Dutch connectives: Change and acquisition as windows on form-function relations*, PhD dissertation Utrecht University. LOT, Utrecht. `https://www.lotpublications.nl/Documents/110_fulltext.pdf`.

Feltracco, Anna, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede (2016), Lico: A lexicon of Italian connectives, *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, Italy, pp. 141–145.

Gao, Qin and Stephan Vogel (2008), Parallel implementations of word alignment tool, *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, Association for Computational Linguistics, Columbus, Ohio, pp. 49–57.

Hoek, Jet (2018), *Making sense of discourse: On discourse segmentation and the linguistic marking of coherence relations*, PhD dissertation Utrecht University. LOT, Utrecht.

Hoek, Jet, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J.M. Sanders (2017), Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study, *Journal of Pragmatics* **121**, pp. 113–131.

Knott, Alistair and Robert Dale (1994), Using linguistic phenomena to motivate a set of coherence relations, *Discourse Processes* **18**, pp. 35–62.

Koehn, Philipp (2005), Europarl: A parallel corpus for statistical machine translation, *MT Summit X: the tenth Machine Translation Summit proceedings*, AAMT, Phuket, Thailand, pp. 79–86. http://mt-archive.info/MTS-2005-Koehn.pdf.

Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan (2014), A PDTB-styled end-to-end discourse parser, *Natural Language Engineering* **20**, pp. 151–184.

Mann, William and Sandra Thompson (1988), Rhetorical structure theory: Towards a functional theory of text organization, *Text* **8**, pp. 243–281.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger (1994), The Penn Treebank: Annotating predicate argument structure, *Proceedings of the Workshop on Human Language Technology*, HLT '94, Association for Computational Linguistics, Plainsboro, NJ, pp. 114–119. https://doi.org/10.3115/1075812.1075835.

Mendes, Amália and Pierre Lejeune (2016), Ldm-pt. a Portuguese lexicon of discourse markers, *Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference*, Budapest, Hungary, pp. 89–92.

Oepen, Stephan, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid (2016), OPT: Oslo–Potsdam–Teesside - Pipelining rules, rankers, and classifier ensembles for shallow discourse parsing, *Proceedings of the CONLL 2016 Shared Task*, Berlin, pp. 20–26.

Pander Maat, Henk (2002), *Tekstanalyse: Wat teksten tot teksten maakt*, Coutinho, Bussum.

Pasch, Renate, Ursula Braußе, Eva Breindl, and Ulrich Hermann Waßner (2003), *Handbuch der deutschen Konnektoren*, Walter de Gruyter, Berlin/New York.

PDTB Research Group (2007), *The Penn Discourse Treebank 2.0 Annotation Manual*. `https://www.seas.upenn.edu/pdtb/PDTBAPI/pdtb-annotation-manual.pdf`.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008), The Penn Discourse Treebank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Redeker, Gisela (1991), Linguistic markers of discourse structure, *Linguistics* **26**, pp. 1139–1172.

Roze, Charlotte, Laurence Danlos, and Philippe Muller (2012), LEXCONN: A French lexicon of discourse connectives, *Discours [En ligne]*. `http://discours.revues.org/8645`.

Sanders, Ted J.M. and Carel van Wijk (1996), PISA – A procedure for analyzing the structure of explanatory texts, *Text-Interdisciplinary Journal for the Study of Discourse* **16** (1), pp. 91–132.

Sanders, Ted J.M., Wilbert P.M.S. Spooren, and Leo G.M. Noordman (1992), Toward a taxonomy of coherence relations, *Discourse Processes* **15** (1), pp. 1–35.

Scheffler, Tatjana, Manfred Stede, Peter Bourgonje, and Felix Dombek (2018), A multilingual database of connectives: connective-lex.info, *Cross-Linguistic discourse annotation: Applications and perspectives*, Toulouse, France, pp. 153–202.

Stede, Manfred (2002), DiMLex: A lexical approach to discourse markers, *in* Lenci, A. and V. Di Tomaso, editors, *Exploring the Lexicon – Theory and Computation*, Edizioni dell'Orso, Alessandria, pp. 1–15.

Stede, Manfred and Arne Neumann (2014), Potsdam Commentary Corpus 2.0: Annotation for discourse research, *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reikjavik, pp. 925–929.

Stede, Manfred and Carla Umbach (1998), DiMLex: A lexicon of discourse markers for text generation and understanding, *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'98)*, Montréal, Canada, pp. 1238–1242.

Tiedemann, Jörg (2012), Parallel data, tools and interfaces in opus, *in* Calzolari, N., K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, pp. 2214–2218.

Van der Vliet, Nynke and Gisela Redeker (2014), Explicit and implicit coherence relations in Dutch texts, *in* Gruber, H. and G. Redeker, editors, *The pragmatics of discourse coherence: Theory and Applications*, Pragmatics & Beyond New Series, Benjamins, Amsterdam, pp. 23–52.

Van Noord, Gertjan (2006), At last parsing is now operational, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles.*, Leuven University Press, Leuven, pp. 20–42.

Van Wijk, Carel and Gerard Kempen (1980), Funktiewoorden: Een inventarisatie voor het Nederlands [An inventory of Dutch function words], *ITL-International Journal of Applied Linguistics* **47** (1), pp. 53–68.

Versley, Yannick (2010), Discovery of ambiguous and unambiguous discourse connectives via annotation projection, *in* Ahrenberg, L., J. Tiedemann, and M. Volk, editors, *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, Northern European Association for Language Technology (NEALT), pp. 83–92.

Wang, Jianxiang and Man Lan (2015), A refined end-to-end discourse parser, *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, Association for Computational Linguistics, Beijing, China, pp. 17–24. http://aclanthology.coli.uni-saarland.de/pdf/K/K15/K15-2002.pdf.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi (2016), A discourse-annotated corpus of conjoined vps, *Proceedings of the 10th Linguistic Annotation Workshop*, Association for Computational Linguistics, Berlin, pp. 22–31.